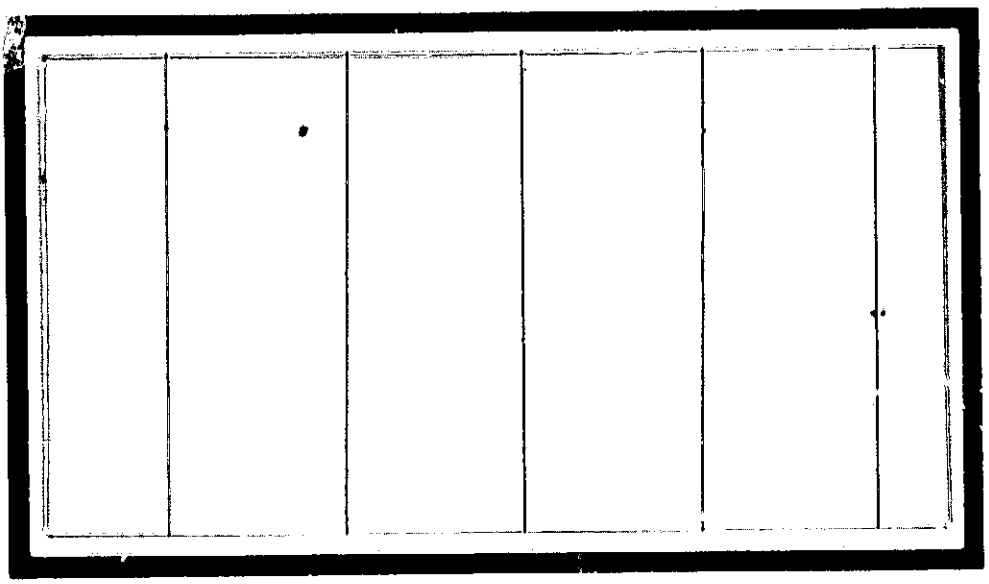

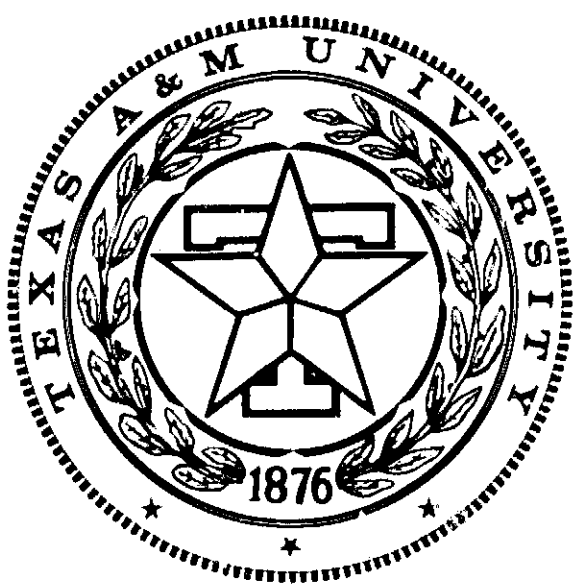

DEPARTMENT OF MATHEMATICS

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS

A METHOD FOR ESTIMATING PROPORTIONS

by

L. F. Guseman, Jr. and Bruce P. Marion

Department of Mathematics
Texas A&M University

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas

Contract NAS-9-13894-1S

April, 1975

Report #4

A METHOD FOR ESTIMATING PROPORTIONS

L. F. Guseman, Jr. and Bruce P. Marion

1. Introduction

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and suppose that $\Omega = \bigcup_{k=1}^{m} \Pi_k$, where each $\Pi_i \in \mathcal{A}$, $\Pi_i \cap \Pi_j = \emptyset$, $i \neq j$, and the unknown a priori probabilities $\alpha_k = P(\Pi_k)$ are positive. Let $X : \Omega \to R^n$ be a random vector with conditional density functions $f_j = f_{X|\Pi_j}$, $1 \leq j \leq m$, and mixture density $f = f_X = \sum_{j=1}^{m} \alpha_j f_j$. Suppose we are given a classification procedure defined by regions R_i, $1 \leq i \leq m$, (which partition R^n) and a decision function c defined for $\omega \in \Omega$ by

$$c(\omega) = i \quad \text{iff} \quad X(\omega) \in R_i .$$

Then the probability that $\omega \in \Omega$ is classified as belonging to Π_i is given by

$$\begin{aligned} P([X \in R_i]) &= P([X \in R_i] \cap (\bigcup_{j=1}^{m} \Pi_j)) \\ &= P(\bigcup_{j=1}^{m} ([X \in R_i] \cap \Pi_j)) \\ &= \sum_{j=1}^{m} P([X \in R_i] \cap \Pi_j) \\ &= \sum_{j=1}^{m} P([X \in R_i] | \Pi_j) \, P(\Pi_j) \\ &= \sum_{j=1}^{m} \alpha_j \, P([X \in R_i] | \Pi_j) \quad . \end{aligned}$$

Let $Y = (Y_1,\ldots,Y_m)^T$ where $Y_i = \chi_{R_i} \circ X$ and χ_{R_j} denotes the characteristic function of the set $R_i \subseteq R^n$, $1 \leq i \leq m$. Then

$$E(Y_i) = E(\chi_{R_i}(X))$$

$$= \int_{R^n} \chi_{R_i}(x)\, f(x)\, dx$$

$$= \int_{R_i} f(x)\, dx$$

$$= \int_{R_i} \sum_{j=1}^{m} \alpha_j\, f_j(x)\, dx$$

$$= \sum_{j=1}^{m} \alpha_j \int_{R_i} f_j(x)\, dx$$

$$= \sum_{j=1}^{m} \alpha_j\, P([X \in R_i] | \Pi_j) \quad .$$

Let $\omega^N = (\omega_1, \omega_2,\ldots,\omega_N)$ be a random sample of size N from Ω. For a given i, $1 \leq i \leq m$, let

$$Y_{i1}(\omega^N) = Y_i(\omega_1)$$

$$Y_{i2}(\omega^N) = Y_i(\omega_2)$$

$$\vdots$$

$$Y_{iN}(\omega^N) = Y_i(\omega_N) \quad .$$

Then for fixed i, $Y_{i1},\ldots,Y_{iN}$ are independent random variables and each has the same distribution as Y_i ([7]); that is, $E(Y_{ik}) = E(Y_i)$, $1 \leq k \leq N$. Letting $\hat{e}_i = \frac{1}{N}\sum_{k=1}^{N} Y_{ik}$, we have $\hat{e}_i(\omega^N) = \frac{N_i}{N}$, where N_i is the number of elements in ω^N that are classified as being from Π_i. If $e_i = E(\hat{e}_i)$, then

$$e_i = E(\hat{e}_i) = E\left(\frac{1}{N}\sum_{k=1}^{N} Y_{ik}\right)$$

$$= \frac{1}{N}\sum_{k=1}^{N} E(Y_{ik}) = \frac{1}{N}\sum_{k=1}^{N} E(Y_i)$$

$$= E(Y_i) = \sum_{j=1}^{m} \alpha_j P([X \in R_i]|\Pi_j) .$$

Letting $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}$, $e = \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix}$, $\hat{e} = \begin{pmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_m \end{pmatrix}$,

we have $e = E(\hat{e}) = P\alpha$, where P is an $m \times m$ matrix whose entry p_{ij}, in the i^{th} row and j^{th} column, is given by

$$p_{ij} = P([X \in R_i]|\Pi_j) = \int_{R_i} f_j(x)\,dx , \quad i,j = 1,2,\ldots,m.$$

We note that a classification procedure produces an estimate $\hat{e}_i(\omega^N) = \frac{N_i}{N}$ of α_i which is biased whenever $e_i = E(\hat{e}_i) \neq \alpha_i$, $1 \leq i \leq m$. The equation $e = E(\hat{e}) = P\alpha$ holds for the error matrix P associated with the

classification procedure used to determine $\hat{e}$ from a given sample. Consequently, an estimate of α could be given by a solution $\hat{\alpha}$ of the following problem:

$$\text{minimize } ||P\alpha - \hat{e}|| \quad \text{(Euclidean norm)}$$

(*)

$$\text{subject to } \sum_{i=1}^{m} \alpha_i = 1, \ \alpha_i \geq 0, \ 1 \leq i \leq m .$$

If P is invertible, then $\bar{\alpha} = P^{-1} \hat{e}$ is an unbiased estimate of α; that is,

$$E(\bar{\alpha}) = E(P^{-1}\hat{e}) = P^{-1} E(\hat{e}) = P^{-1} P\alpha = \alpha .$$

However, simple examples show that even in this case $\bar{\alpha} = P^{-1} \hat{e}$ need not satisfy the nonnegativity constraints even though $\sum_{i=1}^{m} \bar{\alpha}_i = 1$.

For a given P and $\hat{e}$, problem (*) above reduces to the following quadratic programming problem:

minimize the convex functional

(**) $$T(\alpha) = \frac{1}{2} \alpha^T P^T P\alpha - \hat{e}^T P\alpha$$

over the constraint set

$$S = \left\{ \alpha = (\alpha_1, \ldots, \alpha_n)^T : \sum_{i=1}^{m} \alpha_i = 1, \ \alpha_i \geq 0, \ 1 \leq i \leq m \right\} .$$

The functional T is convex (since P^TP is positive semi-definite) and continuous. Since S is compact and convex, a solution always exists. When P is invertible, then P^TP is positive definite and T is strictly convex so that the solution is unique. The above results on convexity of T and uniqueness of the solution can be found in [5].

2. A Method For Computing P and $\hat{e}$

Suppose that each conditional density f_i for the random vector X is multivariate normal with known (or estimated) mean vector μ_i and covariance matrix Σ_i, $1 \leq i \leq m$; that is, $f_i(x) \sim N(\mu_i, \Sigma_i)$, $1 \leq i \leq m$. Under the assumption of equal a priori probabilities (i.e. $\alpha_o = (\frac{1}{m},\ldots,\frac{1}{m})^T$), there exists (see [4]) a $1 \times n$ vector B_o of norm one such that

$$g(B_o) = \min g(B),$$

where

$$g(B) = 1 - \frac{1}{m} \int \max_{1 \leq i \leq m} f_i(y,B)dy$$

and $f_i(y,B) \sim N(B\mu_i, B\Sigma_i B^T)$, $1 \leq i \leq m$. Then the entries in $P = (p_{ij})$ can be readily computed using the expressions

$$p_{ij} = \int_{R_i(B_o)} f_j(y,B_o)dy \ , \ i,j = 1,2,\ldots,m$$

where $R_i(B_o) = \left\{ y \in R^1 : f_i(y,B_o) = \max_{1 \leq j \leq m} f_j(y,B_o) \right\}$, $1 \leq i \leq m$.

Classifying the sample $\omega^N = (\omega_1,\ldots,\omega_N)$ according to the rule

$$c(\omega) = i \quad \text{if and only if } B_o(X(\omega)) \in R_i(B_o)$$

produces the values N_i and hence $\hat{e}_i$, $1 \leq i \leq m$.

The minimizing vector B_o, decision regions $R_i(B_o)$, $1 \leq i \leq m$, and error matrix $P = (p_{ij})$ can be computed using the program LFSPMC described in [3]. The sample $\omega^N = (\omega_1,\ldots,\omega_N)$ is then classified according to the above rule to produce $\hat{e}$ using the classification capability of LFSPMC.

3. Preliminary Numerical Results

The data for the numerical results presented in this section consisted of 30 sets of training statistics and a sample of 16-dimensional vectors of size 8400 obtained from four registered passes (May 5, May 23, June 11, June 29, 1973) of LANDSAT 1 MSS measurements acquired over a 14 square mile test site in Hill County (N), Montana (see [1]). For all runs made the error matrix P was determined from the first of the 30 sets of training statistics provided. A subsample of size 2417 of the original sample was used to compute $\hat{e}$ using the classification procedure which gave rise to P. The sample of size 2417 was made up of vectors from the following five classes: Wheat (784), Fallow (744), Barley (300), Grass (206), and Stubble (383).

Three runs were made using all five classes. Run 1 used LFSPMC and the training statistics from the three registered passes of May 23, June 11, and June 29 to determine P and $\hat{e}$. Run 2 used LFSPMC and the training statistics from the pass of June 11 to determine P and $\hat{e}$. For purpose of comparison, Run 3 used an estimated error matrix determined from a maximum likelihood classification of 12-

dimensional vectors randomly generated using the training statistics for the aforementioned three registered passes. The same classifier was used to determine $\hat{e}$ from the sample of size 2417.

Additional runs were made for the two class case (Wheat, Barley) by using LFSPMC to determine P and $\hat{e}$ from three passes (Run 4) and one pass (Run 5).

For a given P and $\hat{e}$, two quadratic programming algorithms were used to solve problem (**) of the previous section. An algorithm based on the complementary pivot method of Lemke (see [6]) was employed for the case of nonsingular P. In the case where no unique minimum exists (i.e. P singular), a modification of the Frank-Wolfe algorithm [2] due to B. Charles Peters, Jr. was used. The results of the runs are summarized in Tables 1 and 2.

ERROR MATRIX

$$P = \begin{pmatrix} .738 & .003 & .184 & .088 & .018 \\ .003 & .625 & .000 & .145 & .444 \\ .113 & .000 & .809 & .000 & .000 \\ .146 & .206 & .007 & .767 & .192 \\ .000 & .166 & .000 & .000 & .347 \end{pmatrix}$$

CLASSIFIED SAMPLE

$$e = (.288, .264, .137, .189, .121)^T$$

ESTIMATED PROPORTIONS

$$\hat{a} = (.347, .243, .121, .056, .233)^T$$

RUN 1: Five Classes--Three Pass Case
P and $\hat{e}$ Determined By LFSPMC

ERROR MATRIX

$$P = \begin{pmatrix} .752 & .001 & .523 & .052 & .013 \\ .000 & .740 & .004 & .144 & .588 \\ .173 & .000 & .379 & .000 & .000 \\ .075 & .183 & .093 & .805 & .286 \\ .000 & .076 & .000 & .000 & .114 \end{pmatrix}$$

CLASSIFIED SAMPLE

$$\hat{e} = (.332, .357, .098, .183, .030)^T$$

ESTIMATED PROPORTIONS

$$\hat{\alpha} = (.376, .465, .084, .076, .000)$$

RUN 2: Five Classes--One Pass Case
P And $\hat{e}$ Determined by LFSPMC

ERROR MATRIX

$$P = \begin{pmatrix} .965 & .000 & .025 & .005 & .000 \\ .000 & .910 & .000 & .000 & .075 \\ .015 & .015 & .975 & .000 & .000 \\ .010 & .005 & .000 & .970 & .000 \\ .010 & .070 & .000 & .025 & .925 \end{pmatrix}$$

CLASSIFIED SAMPLE

$$\hat{e} = (.316, .271, .142, .080, .192)^T$$

ESTIMATED PROPORTIONS

$$\hat{a} = (.324, .283, .135, .077, .180)^T$$

RUN 3: Five Classes--Three Pass Case
Maximum Likelihood Classifier To Determine $\hat{e}$ And Estimate P

ERROR MATRIX

$$P = \begin{pmatrix} .959 & .038 \\ .041 & .962 \end{pmatrix}$$

CLASSIFIED SAMPLE

$$\hat{e} = \begin{pmatrix} .696 \\ .304 \end{pmatrix}$$

ESTIMATED PROPORTIONS

$$\hat{\alpha} = \begin{pmatrix} .714 \\ .286 \end{pmatrix}$$

RUN 4: Two Classes--Three Pass Case
P And $\hat{e}$ Determined By LFSPMC

ERROR MATRIX

$$P = \begin{pmatrix} .860 & .147 \\ .140 & .853 \end{pmatrix}$$

CLASSIFIED SAMPLE

$$\hat{e} = \begin{pmatrix} .658 \\ .342 \end{pmatrix}$$

ESTIMATED PROPORTIONS

$$\hat{\alpha} = \begin{pmatrix} .716 \\ .284 \end{pmatrix}$$

RUN 5: Two Classes--One Pass Case
P And $\hat{e}$ Determined by LFSPMC

	True Proportions	Estimated P-matrix	Three Pass	One Pass
Wheat	.324	.324	.347	.376
Fallow	.308	.283	.243	.465
Barley	.124	.135	.121	.084
Grass	.085	.077	.056	.076
Stubble	.159	.180	.233	.000

Table 1. Estimated Proportions: Five Classes

	True Proportions	Three Pass	One Pass
Wheat	.723	.714	.716
Barley	.277	.286	.284

Table 2. Estimated Proportions: Two Classes

4. Remarks

The proportion estimation procedure presented in the previous sections has the advantage that the error matrix is determined by the training statistics and thereby requires only one set of ground truth. In addition, the error matrix is the error matrix for the classification procedure used to determine $\hat{e}$. It has the disadvantage that the training statistics must be representative of the mean vectors and covariance matrices for the populations from which the sample was made.

The error matrix is directly related to the probability of misclassification and should be more diagonally dominant with the increase in number of passes used. It should also be mentioned that, under the assumptions of distinct classes and equal a priori probabilities, the error matrix computed by LFSPMC should (barring numerical difficulties) always be nonsingular.

Both of the quadratic programming algorithms used were essentially off-the-shelf programs and require some refinements. The complementary pivot algorithm failed to always meet the problem constraint, $\sum_{i=1}^{m} \hat{\alpha}_i = 1$, to within machine accuracy, and the modified Frank-Wolfe algorithm proved to converge slowly. In any event, the determination of P and $\hat{e}$ using LFSPMC, and subsequent determination of $\hat{\alpha}$ was always accomplished in less than two minutes for the runs reported here. Investigations into the development of more accurate and efficient quadratic programming algorithms are underway.

Theoretical investigations are also underway to extend the feature selection algorithm to the case where the density function for each population is a convex combination of multivariate normal densities. The resulting algorithm gives rise to a method for estimating proportions which involves only two classes; namely wheat and non-wheat.

References

1. W. A. Coberly and P. L. Odell, An empirical comparison of five proportion estimators, Quarterly Progress Report, NASA Contract NAS-9-13512, University of Texas at Dallas, 1975.

2. M. Frank and P. Wolfe, An algorithm for quadratic programming, Naval Research Logistics Quarterly, 3(1956), 95-110.

3. L. F. Guseman, Jr. and Bruce P. Marion, LFSPMC: Linear Feature Selection Program Using the Probability of Misclassification, Report #3, NASA Contract NAS-9-13894, Texas A&M University, Department of Mathematics, January, 1975.

4. L. F. Guseman, Jr., B. Charles Peters, Jr., and Homer F. Walker, On minimizing the probability of misclassification for linear feature selection, Ann. Statist. (To appear).

5. G. Hadley, Nonlinear and Dynamic Programming, Addison-Wesley, Reading, Massachusetts, 1964.

6. A.Ravindran, A computer routine for quadratic and linear programming problems (H), Comm. ACM, 15(1972), 818-820.

7. Howard G. Tucker, An Introduction to Probability and Mathematical Statistics, Academic Press, New York, 1962.